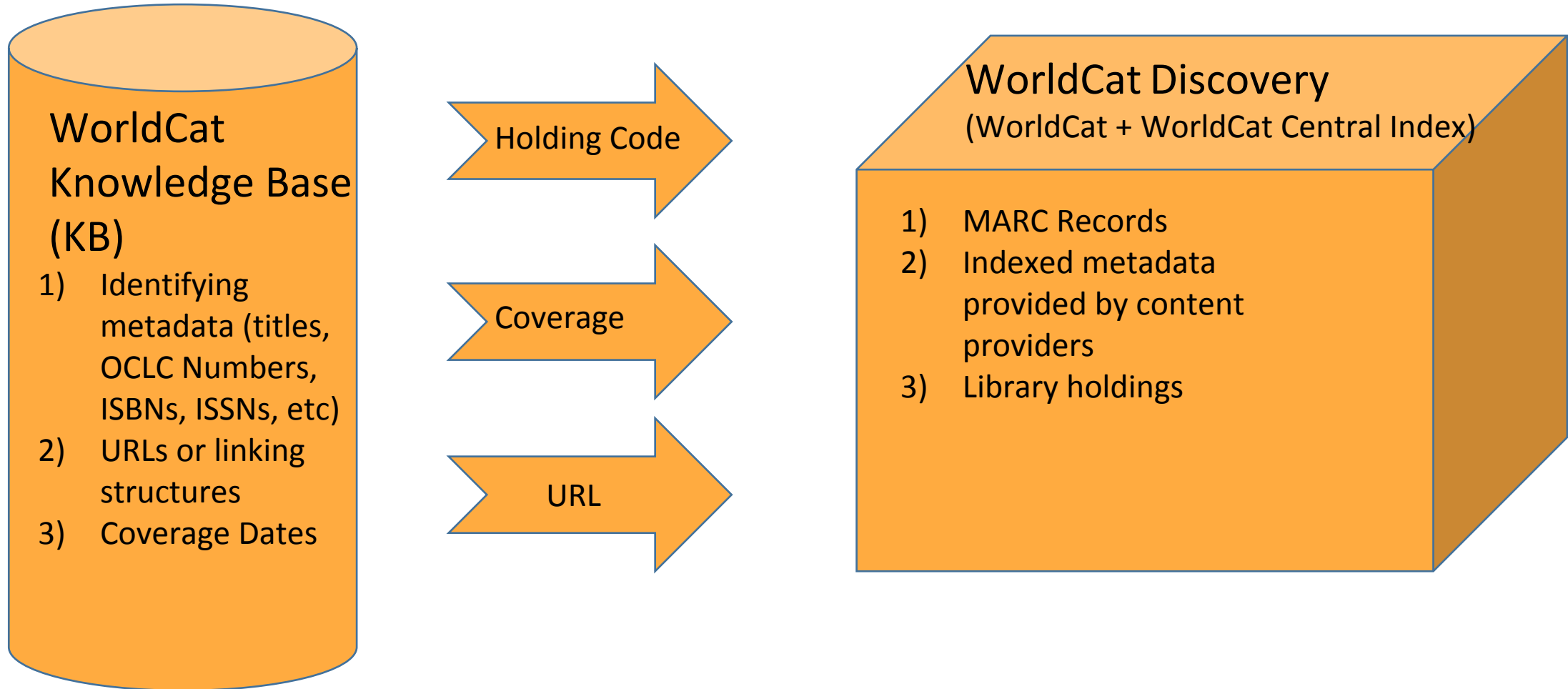


Harvesting MARC Data with the WorldCat Search API

Ben Bradley
Discovery Librarian

Background: Worldcat Discovery and Knowledge Base



KB Data in Discovery

PMLA : publications of the Modern Language Association of America.

 Cite  Link  Email  S



by [Modern Language Association of America.](#)

 eJournal/eMagazine 1888

OCLC Number: 879943264

Held by: [University of Maryland, College Park](#)

▼ Availability

▼ Access Online

[Access journal](#)



Publication

PMLA

Database /

Coverage

Modern Language
Association

(2002~present,
volume:117;issue:1)

▼ Libraries Worldwide

395 Libraries

[Request Item through Interlibrary Loan](#) 

Search location:

McKeldin Library, University of Maryland, Colle



Institution	Symbol	Libraries
University of Maryland, College Park	UMC	UMD Libraries
Howard University Libraries	DHO	
National Academies Research Center	NRZ	
GAO LIBRARY	GAO	GAO; Government Accountability Office General Accounting Office
Executive Office of the President Library	EOP	

What If a Publisher Doesn't Provide Metadata?

Solution

- WorldCat Search API
- Python Script to automate searching and writing records (MARC XML) to a file.
- Use MarcEdit:
 - Convert MARC XML to MARC
 - Export a .tsv file for creating a KBART
- Clean data
- Upload to Collection Manager

How does the script work?

gateway.proquest.com/

The Query

search.proquest.com/

wwwlib.umi.com/



'http://www.worldcat.org/web/services/catalog/search/worldcat/sru?query=
srw.am%3D%22'+quervUrl+'%2a%22'+deg+ebk+lang+'+and+srw.yr%3D'+year+'&
wskey='+wskey+'&servicelevel=full&maximumRecords=100&startRecord=+' +
str(beginPoint)

Handling Large Sets of Results

```
#sends initial request to find the number of results to determine course of action
r = requests.get(url).text
records = r.encode('utf-8')
marcRecords = str(records)
m = re.search(r'<numberOfRecords>(\d+)</numberOfRecords>', marcRecords)
beginpointMax = int(m.group(1))

#if the query returns more than 10,000 results which is greater than can be downloaded, th

if beginpointMax > 10000:
    print('Results greater than 10,000. Performing title searches ' + str(beginpointMax))
    with open(title + ".txt", "a+") as f:
        for letter in alpha:
```


Transformation and Clean-Up

856 42 \$3 ProQuest, Abstract \$u http://gateway.proquest.com/openurl?url_ver=Z39.88-2004&rft_val_fmt=info:ofi/fmt:kev:mtx:dissertation&res_dat=xri:pqm&rft_dat=xri:pqdiss:9994298

856 41 \$3 Texas A&M University \$u <http://lib-ezproxy.tamu.edu:2048/login?url=http://proquest.umi.com/pqdweb?did=727853141&sid=1&Fmt=2&clientId=2945&RQT=309&VName=PQD>

856 41 \$3 Proquest \$u <http://hdl.handle.net/1969.1/Dissertations-2031929>

856 41 \$u <http://wwwlib.umi.com/cr/fullcit?p9992386>

856 42 \$3 ProQuest \$u http://gateway.proquest.com/openurl?url_ver=Z39.88-2004&rft_val_fmt=info:ofi/fmt:kev:mtx:dissertation&res_dat=xri:pqm&rft_dat=xri:pqdiss:9992386

856 41 \$z View Electronic Book (ProQuest Dissertations & Theses Full Text; SDSU users only) \$u <http://search.proquest.com.libproxy.sdsu.edu/docview/304586330?accountid=13758>



Find: .*(/docview/\\d+)\?.* Replace: <https://search.proquest.com>\1

Next Steps

- Script transforming MARC data into KBART
- Update documentation on GitHub
 - <https://github.com/bradley-benjamin26/WCSearchAPIMARCHarvester>
- Find new uses